Toward using bio-ontologies in the Semantic Web: trade-offs between ontology languages

C. Maria Keet and **Mariano Rodríguez**
Faculty of Computer Science, Free University of Bozen-Bolzano, Italy
{keet, mrodriguez}@inf.unibz.it

Abstract

Ontology languages for the Semantic Web have their strengths and weaknesses, in particular in the light of deploying them for biological and medical information systems. We survey and compare the Description Logics-based OWL languages, and the $DL\text{-}Lite$ and $\mathcal{DLR}$ families of languages. Language choices that an ontology developer has to make are, among others, expressivity with n-ary relations (where $n > 2$) and more role properties versus ontology usage for data-intensive tasks. Guidelines are suggested to facilitate choosing the language best fitted for a task.

Introduction

Since the release of the W3C standard of the Semantic Web ontology language OWL in 2004, many bio(medical) ontologies are developed in OWL either *de novo* or have translations from their native language to OWL. An aim is to enhance information integration in biomedical domain and to represent formally our understanding of biological and biomedical reality. However, early-adopters from the bio(medical) domain have already reported their first issues with OWL (Bandini & Mosca 2006; Marshall *et al.* 2006; Ruttenberg, Rees, & Zucker 2006; Smith *et al.* 2006; Wolstencroft, Stevens, & Haarslev 2007). Their problems concern I. (perceived) limitations of ontology languages for representing biomedical knowledge adequately and contain requirements or proposals for improvements of OWL for biomedicine, and II. bottlenecks concerning linking data to the ontologies and subsequent performance issues of the software system when performing common reasoning tasks, such as classification and querying. Applications of biomedical ontologies in the Semantic Web are sparse, but are expected to gain momentum once ontologies can be linked *efficiently* to biological data and used with, *e.g.*, electronic health record management for both annotation and mining hospital information systems, querying whole genomes through an ontology, or even trying to manage the vast amount of metagenomics data (*e.g.*, (Seshadri *et al.* 2007)) through domain ontologies.

Do Semantic Web technologies and ontology languages meet such goals set by domain experts in biology and biomedicine? A familiar requirement is greater expressivity of the ontology language to enable representing the complexities of biology as comprehensive as possible, which corresponds to type I problems mentioned above. This is being addressed gradually, most notably with the recently proposed OWL 1.1[1]. From the computer science perspective, meeting type II requirements, such as data access through an ontology and ontology-based knowledge- or data integration, however, are somewhat 'behind' compared to expectations of science researchers. One tried and tested solution is the Instance Store (Bechhofer, Horrocks, & Turi 2005; Wolstencroft, Stevens, & Haarslev 2007) that links an expressive OWL ontology to a relational database, but is not scalable to large amounts of data or large ontologies – precisely because of the expressive ontology language. A recently proposed alternative is the so-called 'lite' family of ontology languages (Calvanese *et al.* 2006; 2005), which are less expressive but are better scalable—that is, like one is accustomed to from relational databases—and therefore will be more suitable for use with bio-ontologies in large information systems and across the Semantic Web.

To clarify the differences between these new and extant ontology languages and their performance with intended usages, and, more importantly, the unavoidable trade-offs, we compare 9 Description Logics-based ontology languages and provide an overview of the important distinguishing features and limitations in Section 2. Given the identified trade-offs due to expressivity of an ontology language, computational limitations, and (under-)used language features, we suggest guidelines to choose the best suitable formal language for the task at hand (Section 3). Conclusions and ongoing research are described in Section 4. An extended version for this paper with more explanation and examples is available as technical report (Keet & Rodriguez 2007).

Features and limitations of knowledge representation languages

Knowledge representation languages have their origins in logic and a resulting knowledge base system combines the 'model' (logical theory) with data. Knowledge representation languages like Description Logics (DL) are being used as a unifying paradigm for ontology development and formal conceptual modelling (Baader *et al.* 2003). We assess features and limitations of DLs for both biomedical ontologies and conceptual models for biological and medical data.

[1] http://webont.org/owl/1.1/ (Editor's draft of 6-4-2007).

OWL features. Within the scope of the Semantic Web for health care and life sciences, biomedical ontologies, ontology representation languages, and formalisms for biomedical data, the focus is on use of the W3C standard Web Ontology Language OWL. "the OWL language" comes in three flavours: *OWL-full* is built on top of RDF, *OWL-DL* is based on the DL $\mathcal{SHOIN}(\mathcal{D})$, and *OWL-Lite* (a subset of OWL-DL) is based on the DL $\mathcal{SHIF}(\mathcal{D})$. OWL 1.1 is based on the DL language $\mathcal{SROIQ}(\mathcal{D})$ (Horrocks, Kutz, & Sattler 2006), and extends the functionality of OWL-DL with, a.o., several role properties, such as reflexivity and concatenation, and qualified number restrictions that allows for qualified roles (*i.e.*, range defined with a concept). On the other hand, OWL 1.1 functional-style syntax is not backwards compatible with OWL-full, OWL-DL or OWL-Lite abstract syntaxes. The main differences between the DL-based OWL languages are described by (Cuenca Grau *et al.* 2006) and summarised in Table 1.

DL-Lite features. $DL\text{-}Lite$ is a family of DL languages whose expressive power is specifically tailored to provide good performance reasoning algorithms in the presence of large amounts data stored in the ABox ('individuals in the ontology') or linked relational databases (Calvanese *et al.* 2006; 2005). Focusing on ontology-based data access and ontology-based database integration, $DL\text{-}Lite$ allows for delegation of data handling to relational databases through database-ontology mappings and algorithms that translate queries posed in terms of a $DL\text{-}Lite$ ontology to suitable queries over the linked database(s). Modelling features available in the $DL\text{-}Lite$ family—beyond the usual features—are role value-domains and, implicitly, n-ary relations where $n > 2$; see Table 1 for details.

DL languages for formal conceptual modelling. We take a brief look at formal conceptual modelling with DLs, because of the option for common usage of DLs for both ontology and conceptual modelling development, the prospect of ontology-driven information systems, database and tool integration through the use of ontologies, and smoothening translation from an ontology to conceptual models and their corresponding databases. The DL $\mathcal{DLR}$ and its extensions were specifically developed to provide a mapping from conceptual modelling languages such as UML, EER, and ORM2 to a DL (Berardi, Calvanese, & De Giacomo 2005; Calvanese, De Giacomo, & Lenzerini 1999; 1998; Keet 2007) and has a mapping to the DIG interface for DL reasoners, such as RACER and Pellet, to enable automated reasoning over conceptual models. The OWL shortcoming that it cannot deal with "even simple interactions among pluralities of continuants" (Smith *et al.* 2006) is addressed adequately with $\mathcal{DLR}$ that allows for n-ary relations ($n \geq 2$) in the language. $\mathcal{DLR}$s also support primary key identification and functional roles for UML methods (in $\mathcal{DLR}_{\mathsf{ifd}}$), role acyclicity and transitivity, and role concatenation ($\mathcal{DLR}_{\mu}$, $\mathcal{DLR}_{\mathsf{reg}}$), and temporal DL ($\mathcal{DLR}_{\mathcal{US}}$); see Table 1.

Guidelines for choosing the most suitable formal language

The main question is, of course: what do you want to do with the formal ontology or conceptual model? We discuss some common scenarios in this section, and relate them to several extant bio(medical)-ontologies.

Computational limitations and under-used features

The first step in answering the question is to determine what is more important: getting all details correctly represented, *i.e.*, to represent scientific theories as comprehensive as possible, or automated reasoning support (including query answering) over the ontology or conceptual model. The reason for this either-or choice is the direct proportional relation that exists between the computational complexity of reasoning over an ontology and the expressive power of the language used to formalize the ontology. The computational complexity of a problem indicates the rate at which the resources (*i.e.*, computation time and memory) required to solve the problem grow with respect to the size of the problem's input. For instance, the computational complexity of reasoning in OWL-DL is NExpTIME-complete (Cuenca Grau *et al.* 2006) and the $\mathcal{DLR}$ family is in ExpTIME (Calvanese, De Giacomo, & Lenzerini 1998), whereas the $DL\text{-}Lite$ family remains within polynomial time (Calvanese *et al.* 2006). Practically, this means that software systems using OWL 1.1 and $\mathcal{DLR}$-formalized ontologies and conceptual models will grow exponentially slower with every increase in the size of the ontology or the amount of data populating the ontology, whereas systems using $DL\text{-}Lite$ will grow only polynomially, as with relational database systems. Hence, the latter can deal with much larger inputs. For instance, ontologies that are populated by more than a few hundred thousand individuals currently may require hours or days when modelled with and queried through expressive languages instead of the desired seconds or minutes, as observed by, *e.g.*, (Marshall *et al.* 2006) with their HistOn ontology about transcription factor binding sites. Classification of protein phosphatases (Wolstencroft, Stevens, & Haarslev 2007) using the ontology was not scalable either. In some cases, the expressivity of a language might render the reasoning problems computationally undecidable (*e.g.*, OWL-Full), which means that it is impossible to implement systems which provide automated reasoning support for the full language. These inherent limitations cannot be circumvented by experienced software programmers. This might seem a big problem for adoption of Semantic Web technologies by biology and biomedicine, but is not necessarily so.

Identifying necessary *and* sufficient conditions (see "Asserted conditions" in Protégé) for DL's 'defined concepts' rarely occurs in biological and biomedical domain; *e.g.*, the MGED ontology[2] for microarray experiments, mammalian phenotype[3], BioPax level2[4] for biological pathways, and HistOn have only primitive concepts. Put differently, developing a taxonomy tree-only is already quite an achievement, and *the full expressive power of OWL is not used*. Yet, if one has a 'simple' taxonomy or ontology but still uses a reasoner for expressive ontology languages, it uses a range of algorithms for descriptions that could be in the ontology, but are not there. With an ontology that uses a less expressive ontology language, one should be able to take advantage of more efficient reasoning algorithms for the fewer tasks to compute and thereby gain in performance.

[2] http://mged.sourceforge.net/ontologies/MGEDontology.php

[3] http://www.informatics.jax.org/searches/MP_form.shtml

[4] http://www.biopax.org/

Language ⇒ / **Feature** ⇓	OWL			$DL\text{-}Lite$			$\mathcal{DLR}$		
	Lite	**DL**	**v1.1**	$\mathcal{F}$	$\mathcal{R}$	$\mathcal{A}$	*ifd*	μ	*reg*
Role hierarchy (taxonomy of relations)	+	+	+	-	+	+	+	+	+
N-ary roles (where $n \geq 2$, ternary, quaternary relation etc.)	-	-	-	$\pm$	$\pm$	$\pm$	+	+	+
Role concatenation (limited role composition)	-	-	+	-	-	-	-	-	+
Role acyclicity (least fixpoint construct)	-	-	-	-	-	-	-	+	-
Symmetry	+	+	+	-	+	+	-	-	-
Role values (role attribute values, like strings and integers)	-	-	-	-	-	+	-	-	-
Qualified number restrictions	-	-	+	-	-	-	+	+	+
One-of, enumerated classes	-	+	+	-	-	-	-	-	-
Functional dependency (or UML method)	+	+	+	+	-	+	+	-	+
Covering constraint over concepts (total/complete covering)	-	+	+	-	-	-	+	+	+
Complement of concepts (disjointness of classes)	-	+	+	+	+	+	+	+	+
Complement of roles (disjointness of roles)	-	-	+	+	+	+	+	+	+
Concept identification (primary key with $>$ attribute)	-	-	-	-	-	-	+	-	-
Range typing (define concept of the 2nd participant in role)	-	+	+	-	+	+	+	+	+
Reflexivity *	-	-	+	-	-	-	-	+	+
Antisymmetry *	-	-	-	-	-	-	-	-	-
Transitivity * ‡	+	+	+	-	-	-	-	+	+
Asymmetry ‡	+	+	+	-	+	+	-	$\pm$	-
Irreflexivity ‡	-	-	+	-	-	-	-	+	-

Table 1: Differences between DL-based ontology and conceptual modelling languages; terms in braces are regularly considered as synonyms; indirect or implied support ($\pm$); properties of the parthood (*) and proper parthood (‡) relation.

Ontology	**Characterizing DL**[5,6]
ProPreO[7]	$\mathcal{SHOIN}(\mathcal{D})$
BioPAX	$\mathcal{ALCHON}(\mathcal{D})$
Cell Cycle Ontology	$\mathcal{SIN}(\mathcal{D})$
HistOn	$\mathcal{ALCHIF}(\mathcal{D})$
NMR Ontology[8]	$\mathcal{SHF}$
MGED Ontology	$\mathcal{ALEOF}(\mathcal{D})$
Gene Ontology	$\mathcal{ALE}(\mathcal{D})$
Protein-Protein Interaction	$\mathcal{ALE}(\mathcal{D})$
Mammalian Phenotype	$\mathcal{AL}(\mathcal{D})$
FungalWeb[9]	$\mathcal{FL}_0$

Table 2: DL characterization of the expressivity of several bio-ontologies sorted in (approximate) decreasing order with respect to the complexity of the language.

We illustrate this briefly for several bio-ontologies.

Example. Similarly to the way that OWL 1.1, OWL-DL and OWL-Lite are characterized by a DL, the expressivity used in an ontology represented with OWL is also characterized by a DL which can be identified by analysing the language constructs used in it. We present such an analysis for the previously mentioned ontologies and some other well known bio-ontologies in Table 2. Given the languages and analysis of the examined ontologies, we can see that the current Gene Ontology taxonomies, Protein-Protein Interaction ontology[10], and HistOn, among others, remain within $DL\text{-}Lite_{\mathcal{A}}$ expressivity. The BioPax and MGED ontologies can be adapted easily to match $DL\text{-}Lite_{\mathcal{A}}$ by correcting the misguided modeling of the ontology's versioning information using the `oneOf` construct instead of OWL's annotation facilities. On the other hand, the developers of the Foundational Model of Anatomy and ProPreO ontologies aim to be as comprehensive as possible and therefore use almost the full power of OWL-DL. Subsequently, one may be able to extract a 'light' version of an ontology that fits into $DL\text{-}Lite$ expressivity to aid implementation of *e.g.*, database integration in the biomedical domain. $\diamondsuit$

Ontology language choices

Based on the analysis of language features, computational limitations, and (under-)usage of language features, we propose several guidelines to choose the (relatively) optimal ontology language for the intended core tasks.

I. Comprehensiveness

a. *No computation*. The user can choose freely the language that covers to the best extent the expressive requirements of the ontology. Suitable languages are OWL 1.1 and the $\mathcal{DLR}$ family, or to resort to other logics that are currently largely outside of the scope of the Semantic Web, such as first- or higher order logics, temporal logics, epistemic logic etc.. For instance, to represent a scientific theory as comprehensive as possible and for foundational ontologies, such as BFO[11].

[5] See (Baader *et al.* 2003) for an overview of the DLs presented. Results were obtained with Protégé's and SWOOP's *DL expressivity metric* facilities

[6] Sample date: 12-2-2007.

[7] http://lsdis.cs.uga.edu/projects/glycomics/propreo/

[8] http://obo.sourceforge.net/cgi-bin/detail.cgi?nmr

[9] http://www.cs.concordia.ca/FungalWeb/

[10] http://psidev.sourceforge.net/mi/xml/doc/user/index.html

[11] http://www.ifomis.uni-saarland.de/bfo/home.php

b. *Computation is desired and plenty of time and memory is available*. A decidable language has to be used. The size of the ontology or the data will be limited by the resources at hand, that is: either a large ontology of universals or a small one that can be linked to a small amount of data. Languages suitable for this setting are OWL-DL, OWL 1.1, the $\mathcal{DLR}$ family. For instance, developing integrated conceptual models that are used 'off-line' for eventual data integration and developing reference ontologies, such as the FMA, for particular subject domains.

II. Computation

a. *Computing time and memory are an important component*. This is a grey area as to what constitutes a reasonable amount of waiting time, and either OWL-DL, OWL-Lite or $DL\text{-}Lite$ could be used: the former two if there is relatively little data (with as rule-of-thumb, certainly less than hundred thousand instances) and if the ontology is small (less than a few hundred DL-concepts); $DL\text{-}Lite$ can be used in all scenarios.

b. *Computing time and memory are critical*. The accuracy of the ontology will be limited compared to item I, but its size and the amount of data linked to the ontology can be as for relational databases. Languages suitable for this setting are those in the $DL\text{-}Lite$ family. For instance, to pose complex queries over the data, like microarray data and data about large genomes, through ontologies such as the GO, MGED ontology, and HistOn.

With these main four distinctions, one could construct a decision procedure, as in "if you want an ontology to do x, then...". However, the four distinctions remain and can be reused for any new scenario, whereas a decision tree would have to be updated upon each usage variation.

Conclusions

Based on, and motivated by, a comparative assessment of ontology- and formal conceptual modelling languages, current bio-ontologies and their usage, and prospective scenarios for ontology-based and ontology-mediated tasks, we provided guidelines for choosing the optimal ontology language for the task. Although it is expected that ontology languages develop further, the main trade-off between expressivity and usability in data-intensive biomedical information systems remains.

Current research comprises mapping $DL\text{-}Lite$ to OWL 1.1 and incorporating a DIG API for the QuOnto system[12], which will enable easy adoption of the $DL\text{-}Lite$ languages by current OWL/Protégé users. We plan to conduct a more comprehensive analysis of the (under-)used ontology language capabilities, develop algorithms for 'lite'-izing expressive ontologies to use for ontology-based data access and integration, and reasoning services for bio-ontologies.

Acknowledgments. The authors thank Diego Calvanese for helpful suggestions on an earlier draft.

References

Baader, F.; Calvanese, D.; McGuinness, D.; Nardi, D.; and Patel-Schneider, P. F., eds. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.

Bandini, S., and Mosca, A. 2006. Mereological knowledge representation for the chemical formulation. In *Proc. of FOMI2006*, 55–69.

Bechhofer, S.; Horrocks, I.; and Turi, D. 2005. The OWL Instance Store: System description. In *Proc. of CADE-20*, LNCS. Springer Verlag.

Berardi, D.; Calvanese, D.; and De Giacomo, G. 2005. Reasoning on UML class diagrams. *Artificial Intelligence* 168(1–2):70–118.

Calvanese, D.; De Giacomo, G.; Lembo, D.; Lenzerini, M.; and Rosati, R. 2005. DL-Lite: Tractable description logics for ontologies. In *Proc. of AAAI 2005*, 602–607.

Calvanese, D.; De Giacomo, G.; Lembo, D.; Lenzerini, M.; Poggi, A.; and Rosati, R. 2006. Linking data to ontologies: The description logic *dl-lite*$_{\mathrm{a}}$. In *Proc. of OWLED 2006*.

Calvanese, D.; De Giacomo, G.; and Lenzerini, M. 1998. On the decidability of query containment under constraints. In *Proc. of PODS 1998*, 149–158.

Calvanese, D.; De Giacomo, G.; and Lenzerini, M. 1999. Reasoning in expressive description logics with fixpoints based on automata on infinite trees. In *Proc. of IJCAI99*, 84–89.

Cuenca Grau, B.; Horrocks, I.; Parsia, B.; Patel-Schneider, P.; and Sattler, U. 2006. Next steps for OWL. In *Proc. of OWLED-2006*.

Horrocks, I.; Kutz, O.; and Sattler, U. 2006. The even more irresistible $\mathcal{SROIQ}$. *Proc. of KR-2006* 452–457.

Keet, C. M., and Rodriguez, M. 2007. Comprehensiveness versus scalability: guidelines for choosing an appropriate knowledge representation language for bio-ontologies. Technical Report KRDB07-5, Faculty of Computer Science, Free University of Bozen-Bolzano. http://www.inf.unibz.it/krdb/pub/.

Keet, C. M. 2007. Prospects for and issues with mapping the Object-Role Modeling language into DLRifd. In *Proc. of DL'07*. CEUR-WS.

Marshall, M. S.; Post, L.; Roos, M.; and Breit, T. M. 2006. Using semantic web tools to integrate experimental measurement data on our own terms. In *Proc. of KSinBIT'06*, volume 4277 of *LNCS*, 679–688. Springer Verlag.

Ruttenberg, A.; Rees, J.; and Zucker, J. 2006. What biopax communicates and how to extend owl to help it. In *Proc. of OWLED 2006*.

Seshadri, R.; Kravitz, S.; Smarr, L.; Gilna, P.; and Frazier, M. 2007. CAMERA: A community resource for metagenomics. *PLoS Biology* 5(3):e75.

Smith, B.; Kusnierczyk, W.; Schober, D.; and Ceusters, W. 2006. Towards a reference terminology for ontology research and development in the biomedical domain. In *Proc. of KR-MED 2006*.

Wolstencroft, K.; Stevens, R.; and Haarslev, V. 2007. Applying owl reasoning to genomic data. In Baker, C., and Cheung, H., eds., *Semantic Web: revolutionizing knowledge discovery in the life sciences*. Springer: NY. 225–248.

[12] http://www.dis.uniroma1.it/∼quonto